DOCUMENT RESUME

ED 230 580 TM 830 293

AUTHOR Schaefer, Mary M.; Gross, Susan K.
TITLE A Comparison of Reliability Estimates from Single and Double Administrations of Criterion-Referenced Tests.
PUB DATE Apr 83
NOTE 40p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; Correlation; *Criterion Referenced Tests; Cutting Scores; Elementary Secondary Education; Estimation (Mathematics); Instructional Systems; *Mathematical Models; Norm Referenced Tests; *Pretests Posttests; *Test Reliability
IDENTIFIERS Test Length

ABSTRACT

Viewing the reliability for criterion-referenced tests as that of mastery classification decisions, three models for determining reliability were examined using two test administrations so that two estimates could be compared to a standard. A major purpose of the research was to determine how several reliability coefficients (coefficient kappa, an estimate of kappa, and a coefficient of agreement) were affected by test length, test length and cutoff score combined, student ability and cutoff score combined, sample size, and test content heterogeneity. The relationships between reliability and validity for criterion-referenced tests, and between norm-referenced and criterion-referenced reliability for the same data, were also examined using the Instructional System in Mathematics (ISM) program, an objectives-based instructional management system for grades K-8. A sample of 325 students were tested with three criterion-referenced tests before and after a two-week spring vacation. Coefficient kappa, which can only be calculated when there are two test administrations, generally had the lowest values and the largest standard errors. The estimate of coefficient kappa had slightly larger mean values and smaller standard errors. The coefficient of agreement was the highest reliability coefficient. (CM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED230580

A COMPARISON OF RELIABILITY ESTIMATES FROM SINGLE AND DOUBLE ADMINISTRATIONS OF CRITERION-REFERENCED TESTS

Mary M. Schaefer*

Susan K. Gross

Montgomery County Public Schools
Rockville, MD 20852

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

M.M. Schaefer

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Printed in the U.S.A

Current address: The Center for the Study of Testing, Evaluation, and Educational Policy

Boston College
Chestnut Hill, MA 01760

TM 830 293

One of the problems inherent in the increased use of criterion-referenced tests for making decisions about educational progress has been the problem of determining the reliability of the tests. The determination of reliability for criterion-referenced tests has been conceptualized in several different ways.

The first concept, the reliability of mastery classification decisions, is concerned with the consistency with which individuals are classified as masters or nonmasters on two forms of a criterion-referenced test. Masters are individuals who meet or exceed the standard of performance (or criterion) while nonmasters are those whose achievement does not equal the criterion level of performance.

The second type of reliability, the reliability of criterion-referenced test scores, refers to the stability of deviations from the criterion score on two forms of a criterion-referenced test. This concept of reliability is concerned with the stability of a student's score deviation from the cut-off score on two administrations of a criterion-referenced test.

The third type of reliability, the reliability of domain score estimates, is different from the first two types in that it does not involve setting a criterion score and thus does not determine masters and nonmasters. This reliability concerns the consistency of students' scores across two forms

of parallel tests.

For more complete descriptions of these concepts and of the various techniques for determining the reliability of criterion-referenced tests the reader is referred to Hambleton, Swaminathan, Algina and Coulson (1978) and to Berk (1980).

Since much of the use of criterion-referenced tests has been for the purposes of deciding on promotion in schools either from grade to grade or into another level or type of subject matter, the reliability of mastery classification decisions has been the focus of several conceptualizations of reliability for criterion-referenced tests. In subscribing to this view, one is concerned with the reliability of the decision that a student is a master or a nonmaster of material such that if a student were tested a second time would the mastery decision remain the same. The difficulty of course is that the use of criterion-referenced tests in instances of this type of decision-making often precludes the administration of the criterion-referenced test a second time.

In focusing on the view of reliability for criterion-referenced tests as the reliability of mastery classification decisions, three models for determining reliability have been proposed in the past several years: coefficient kappa, k, (Cohen, 1960) suggested for use with criterion-referenced tests by Swaminathan, Hambleton, and

Algina, (1974), an estimate of coefficient kappa, k(H) (Huynh 1976) and the coefficient of agreement, pc(s) (Subkoviak, 1976). Coefficient k is based on two administrations of a criterion-referenced test while the estimate of kappa and the coefficient of agreement were developed for situations where only data from one test administration were available. When data are available from two test administrations one can get an empirical measure of reliability, kappa, but the availability of data from only one administration of a test is a common occurrence. Therefore the purpose of this research was to examine these three models for reliability in a situation where data were available from two test administrations so that the two estimates could be compared to a standard. Student outcomes on criterion-referenced tests were examined in a test-retest setting to determine the extent of test decision variability under certain conditions and the sensitivity of the three estimates to these conditions.

The major purposes of the research may be summarized as follows:

1. How is coefficient kappa, k, the standard, affected by test length, test length and cut-off score combined, student ability and cut-off score combined, sample size and test content heterogeneity?
2. How are k(H) and pc(s), the estimates of reliability

affected by the aforementioned variables?

3. Is k(H) a good estimate of k for these differeng conditions?
4. What is the relationship between reliability and and validity for criterion-referenced tests?
5. What is the relationship between norm-referenced and criterion-referenced reliability for the same data?

The data from the present research came from the Instructional System in Mathematics (ISM) program in the Montgomery County (Md.) public schools, an objectives-based instructional management system for grades K-8. In the ISM program, the students are tested with criterion-referenced tests at the completion of work in each objective. The objectives are arranged in hierarchical sequence within each skill area. The sample consisted of 325 students who were tested with three criterion-referenced (ISM) tests before and after spring vacation, resulting in a two-week interval between test administrations.

To investigate the effect of test length on the coefficients, the data from one test were reorganized to construct tests of different lengths. To control for possible variation in item difficulty all possible combinations of items for a test of a given length were constructed from the pool of items measuring one objective. If n=total number of items and p=chosen test length, the total number of

combinations is given as comb= $_nC_p = \frac{n!}{p!(n-p)!}$. For each test length for each objective (5 objectives) all three reliability coefficients were computed. Rank order correlations were computed between k and each of the two estimates.

To examine the combined effect of test length and cut-off score, the three reliability coefficients were computed for all possible combinations of items for a test of a given length for cut-off scores of 60, 80 and 100 percent. Rank order correlations were computed between k and each estimate by cut-off score and test length.

To examine the combined effect of ability level and cut-off score, scores from the Iowa Test of Basic Skills (ITBS) were used to identify students of high and low ability. The analysis of cut-off score and test length was then run in each ability group and the reliability coefficients and rank order correlations were computed as described above.

To examine the effect of sample size the combined effect of cut-off score and test length analysis was run in samples of one (N=25) and two (N=50) classrooms. All three reliability coefficients were computed for each test length and cut-off score in each class size and rank order correlations were computed.

An examination of test content heterogeneity was undertaken by controlling for mastery level and manipulating the types of items on tests of different lengths. Items were

selected at random from different objective levels of the tests. Each reliability coefficient was computed for each of 100 randomly generated item combinations for diffferent test lengths. Rank order correlations were computed over test length.

Additionally, in order to investigate whether k(H) was a good estimate of k for differing conditions, mean difference scores between k and k(H) were computed for each of the separate analyses of test length, test length and cut-off score, ability level and cut-off score, sample size and test content heterogeneity.

This analysis was not undertaken for the coefficient of agreement because it is not intended to be an estimate of K. The scaling for pc(s) is different from that for k and k(H) The coefficient pc(s) does not correct for chance agreement i.e., agreement due to the proportion of masters and nonmasters in the group and thus is indicative of the total proportion of consistent mastery classifications which occur on two tests. Coefficient k and k(H) n the other hand take out the part of the consistency which is due to the mastery/nonmastery composition of the group taking the test (pc). Coefficients k and kH thus can be considered measures of test consistency while pc(s) is a measure of test consistency as well as consistency resulting from the proportion of masters and nonmasters in the group tested.

Given this fact the size of the pc(s) coefficient will always (theoretically) be larger than the k or k(H) coefficients.

To address the issue of predictive validity, students' scores on the ISM tests were compared to their scores on the ITBS. Finally, norm-referenced coefficients were computed for cut-off score and test length and were correlated with the criterion-referenced coefficients.

I. Test Length

The results of the test length analysis showed that there was an increase in the mean values of all the reliability coefficients up to a test length of 8 items. Beyond that the results were unpredictable and there was no clear trend. Coefficient k was lowest for the longer test lengths and an increase of the other coefficients was difficult to detect due to the size of the standard errors. (see Table 1) Rank order correlation coefficients between k and the other coefficients were all below .40, indicative of little agreement among the coefficients.

The reason for the longer tests having the lower mean values for reliability may be that these tests were taken by the younger students. An examination of the data on these items indicates that for the 13-item test taken by the third graders, 23 percent of the students' scores changed from mastery to nonmastery or from nonmastery to mastery between

the two test administrations. Those items with the most inconsistent performance were the last three items on the test, suggesting that fatigue and/or boredom might have been responsible for the inconsistent performance.

The results of the difference score analysis indicated that k(H) is only a good estimate of k for the shorter test lengths. Additionally k(H) was consistently higher than k, particularly at the longer test lengths.

Similar findings about the size of the coefficients and their standard errors for different test lengths have been reported by Subkoviak (1978). Specifically, Subkoviak (1978) also reported that coefficient, k, produced relatively large errors of estimation for classroom size samples and that estimates of pc(s) tended to be overestimates of parameter values when the cut-off score was in the tails of the distribution, such as 80 percent.

With respect to k(H), however, Subkoviak (1978) reported that Huynh's (1976) method produced underestimates for all criterion levels for a 10-item test. In this research k(H) was generally consistently higher than the standard, k.

II. The Combined Effect of Cut-Off Score and Test Length

The three reliability coefficients behaved differently as the cut-off score was raised. The three cut-off scores

examined were 60, 80, and 100 percent. Most of the maximum mean values of k(H) occurred at the 80 percent cut-off score and there was a considerable decrease at the 100 percent cut-off score. The coefficients of agreement, and k, however, were at their maximum mean value at the 60 percent cut-off score and the mean values declined steadily to the 100 percent cut-off score. Thus for k(H) the results took the form of an inverted u and the results for k and for pc(s) (simple and compound) were a decreasing straight line. (see TABLE 2)

Interestingly all the coefficients tend to be lowest when the cut-off score is set at 100 percent, suggesting that setting a cut-off score so high would result in very low reliability.

The results for k(H) would support those of Huynh (1976) who found that the value of k(H) increases as the cut-off score increases, up to a point and then it begins to decline. When the cut-off score is too low or too high (such as 100 percent) pc is near one and this may explain why k(H) declines after reaching a maximum value.

Subkoviak (1978) reported that the procedure for computing the coefficient of agreement seemed to produce slight underestimates when the cut-off score was in the center of the test score distribution such as 50 percent and slight overestimates when the cut-off score was in the tails of the distrbution such as 80 percent. In this research the highest

values of pc(s) occurred when the cut-off score was toward the center of the distribution (e.g., 60 percent) and the values decreased as the cut-off score was raised.

The only strongly positive relationships among the coefficients by test length occur at the 60 percent cut-off score. There is virtually no agreement between the coefficents at either the 80 or 100 percent cut-off score, despite the fact that all the coefficients reach their lowest mean values at the 100 percent cut-off score.

The use of k(H) as an estimate of k based on the difference score analysis appears reasonable only for the shorter test lengths using cut-off scores of 60 and 80 percent. However, even when k(H) appears close in value to k, it is always consistently higher than k which may indicate that it overestimates k at all cut-off scores.

III. The Combined Effect of Ability Level and Cut-Off Score

There was little consistent change in the size of the coefficients as the cut-off score was raised for either of the ability groups. Consistent with the analysis of test length and cut-off score the coefficients are at their lowest at the 100 percent cut-off score for both ability groups. Also consistent with previous analyses is the large standard errors for k. (see Table 3)

With respect to differences in the sizes of the coefficients for the two ability groups, both k and kH are slightly larger for the low ability group. An examination of the data reveals that the low ability group had more consistent decisions -- consistent nonmastery -- than the high ability group which would result in higher values for k.

Neither of the pc(s) coefficients shows any consistency in being higher for one ability group or the other. Rank order correlation coefficients computed between k and the pc(s) coffiicients in the two ability groups shows no relationship between the two coefficients for the low ability group while the correlations between pc(s) (simple) and k are positive at the 60 percent and 80 percent cut-off scores for the high ability group. The low ability group is more heterogeneous with respect to math achievement scores on the ITBS with scores which range from the 2nd percentile to the 49th percentile. The high ability group, however, has scores which cluster between the 75th and 98th percentiles. The difference in the homogeneity of the groups may be the difference in the relationships between the pc(s) coefficients and k.

Subkoviak (1978) stated that if a group is composed entirely or almost entirely of high ability students that fact alone would assure a large proportion of consistent mastery/mastery outcomes, and thus a high value for pc(s) The same should hold true for a group of low ability students --

with a large proportion of consistent mastery/nonmastery outcomes pc(s) should also be high. In this research neither of the pc(s) coefficients was consistently higher for the low or the high ability group. Rather for each cut-off score of 60, 80 or 100 percent the group with the highest proportion of consistent decisions has the highest mean values for pc(s). An examination of the data reveals that higher mean values for the high ability group reflect a high proportion of students mastering both tests. Similarly higher mean pc(s) values for the low ability group at the 100 percent cut-off is indicative of a high proportion of students not mastering either test.

Two analyses investigated the relationship between k and k(H). The rank order correlation coefficients computed between k and k(H) differ by ability group. For the low ability group, there is a positive correlation between k and k(H) only at the 100 percent cut-off score.

The positive correlation between k and k(H) at the 100 percent cut-off score is perhaps indicative of low reliability for this cut-off score, regardless of the homogeneity or heterogeneity of the groups. All the reliability coefficients were lowest at the 100 percent cut-off and the positive relationship between k and k(H) is possibly a reflection of that phenomenon.

There is a positive relationship between k and k(H) for the high ability group -- at the 60 percent and 80 percent

cut-off scores. The positive correlations may be indicative of the high ability group being more homogeneous than the low ability group and thus resulting in k(H) being a better estimate of k for this group. It is suspected that the 100 percent cut-off score results in such inconsistent performance that even for a high abilitygroup where one would not expect k to be so low, k(H) is simply a poor estimate of k.

An analysis of the mean difference scores between k and k(H) resulted in k(H) appearing to be a better estimate of k for the high ability group, at least at the 60 percent cut-off score. There is not much difference between k(H) as an estimate of k for either ability groups for the 80 percent cut-off score and k(H) is a poor estimate of k for both ability groups at the 100 percent cut-off score.

Huynh (1976) reports that test score variability has a positive relationship to k(H). This is supported by this research in that the low ability group for whom the mean values of k(H) were higher has been found to be more heterogeneous in math achievement than the high ability group. The heterogeneity of ability then seems to increase the value of k(H) yet it results in k(H) being a poor estimate of k.

IV. Sample Size

The value of k show no predictable behavior from one sample size to another for the different objectives for which

data were analyzed. Overall the mean values of k are slightly higher for the two classroom samples (n = 50). The size of the standard errors were always larger for the smaller samples. It should be noted, however, that even though the standard errors are smaller in the two classroom sample, they are still large (.072-.089) and quite a bit larger than the standard errors for the k(H) and pc(s) coefficients. (see Table 4)

The k(H) coefficients are somewhat larger in the two classroom sample and the size of the standard errors is smaller in the two classroom sample. However, the standard errors for both one (n = 25) and two classroom (n = 50) samples across all grade levels, ranged from .022 to .110 for one classroom samples and from .022 to .093 for the two classroom samples.

Subkoviak reported that Huynh's (1976) procedure resulted in estimates with relatively small standard errors for classroom size samples. Since Subkoviak's (1978) finding was based on data from the SATs it is possible that the items from the ISM tests are more unreliable and thus the standard errors are larger in estimating the reliabiility.

Overall the analysis of effect of sample size on the coefficients of agreement seems to indicate that the reliability coefficients for the one classroom samples were higher than those for the two classroom samples, at least at

the 60 and 80 percent cut-off scores. The size of the standard errors of both the pc(s) coefficients changes very little from one classroom to two classrooms. The standard errors are very small (.010 and below) for both groups.

There were changes in the rank order correlations between k and the other coefficients between the one and two classroom size samples only at the 60 percent cut-off score. There are strong positive correlations between k and k(H) for both the one and two classroom samples at the 60 percent cut-off score, while there was no relationship between them at any other cut-off score in either sample size. In the two classroom samples there was a moderate correlation, between k and the coefficients of agreement at the 60 percent cut-off, but there was no agreement between them at any other cut-off score for either sample size.

For the shorter test lengths with a cut-off score of 60 percent k(H) appears to be a fairly good estimate of k for the two classroom samples although the estimates are somewhat higher than k. At the 80 percent and 100 percent cut-off scores the mean difference scores between the two coefficients are large for both sample sizes. Additionally k(H) is consistently higher than k for all test lengths at these two cut-off scores.

V. Item Heterogeneity

The necessity for having tests or at least groups of items measuring only one skill has been recommended in the literature on criterion-referenced testing for some time. Thus, one might consider that the determination of reliabilty for a heterogeneous test might lead to lower estimates of reliability. The results of the analysis of the effect of heterogeneous test items on k revealed that for the multiplication items for several groups the mean values for k were higher for the heterogeneous items then for the homogeneous items. (see TABLE 5)

Regardless of level of the students, k does not appear to be sensitive to the inclusion of items measuring different skills on one test, at least with respect to the multiplication items. Perhaps the items which are purportedly measuring different multiplication skills are in fact not measuring different skills, resulting in higher mean values for k for supposedly heterogeneous items.

For all the samples which took homogeneous and heterogeneous tests of division items, the mean values of k were lower for the heterogeneous items. The division items in the pool cover a wider range of objective levels than the multiplication items did and, therefore, it is more likely that more of the students had not mastered the skills measured by some of the items. Thus for the division items, construction of tests with items measuring different skills

seems to have resulted in lower mean values for k.

Since the estimate of k(H) is based on the assumption that the distribtuion of the test score is binomial, (one assumption of which is that the probability of a correct response is constant for every item), it had been predicted that construction of a test with items which vary in content would have a substantial effect on k(H). This is in fact the case since in every analysis the mean values of k(H) were a great deal lower for the heterogeneous items than for the homogenecus items.

It had been predicted that the violation of test homogeneity would have a substantial effect on the coefficient of agreement for the same reason that it affected k(H). The heterogeneous content of the tests did seem to affect the coefficients of agreement. but in the opposite way than expected. Rather than decreasing the size of the coefficients in most cases the violation of test homogeneity increased the size of the coefficients. While the heterogeneity of test content seemed to decrease the precision of estimation for pcs(s), it appeared to increase the precision of estimation for pcs(c).

There is no agreement based on rank order correlation between k and the pc(s) coefficients for heterogeneous items. There also was no agreement between the pc(s) coefficients and k

for tests of homogeneous items, using this cut-off score.

Since the inclusion of heterogeneous items does violate the assumptions of a binomial model, one would expect k(H) to be a poor estimate of k for these tests. However, the rank order correlation computed between k and k(H) for the heterogeneous items was .944. This can be contrasted with the .315 correlation between k and k(H) when the tests were homogeneous.

The stronger correlational relationship between k and k(H) which exists for the heterogeneous items is substantiated by the mean difference scores between k and k(H) for these items. For the heterogeneous items, k(H) is a much better estimate of k than it is for the homogeneous items and this is the only analysis for which k(H) is not consistently higher than k. in fact k(H) is a better estimate of k for these items than for any other analysis.

One aspect of the tests examined in this research which must be considered, particularly in light of the results of the analyses with heterogeneous items, is the construction of the tests. If the test items were less homogeneous by objective or less heterogeneous in the mixing of objectives, it is possible that the particular combinations of items selected in the random selection of items were responsible for the inconsistent results not only among the coefficients but also within the analyses for each individual coefficient. While a random sample of 100 combinations should be adequate

to reduce bias due to selection of items, another explanation for these erratic results does not see apparent.

VI. Validation

For both groups of students, both the reliability and the validity coefficient are very low (all below .50). This would suggest that the multiplication items taken by the students may not be particularly good measures of the skills they are intended to measure. Neither do the items reveal high performance consistency nor do they reveal a high degree of relationship to an outside criterion of performance. Thus, the validity was low. It does, however, seem to be the case for the criterion-referenced tests that validity is bounded by the square root of reliability as is the case for norm-refernced tests.

Overall the criterion-referenced test items in the two multiplication and two division mastery tests given by MCPS and analyzed here do not appear to have much validity in terms of an outside criterion of proficiency. The only group of items which has an even moderate relationship to the ITBS was one mastery test in Division 10-N and this relationship was not that strong (r = .438). (see TABLE 6)

VII. Relationship Between Norm-Referenced and Criterion-Referenced Coefficients

The relationship between the criterion-referenced coefficients and the KR 21 and test-retest coefficients was examined by means of computing Pearson Product Moment correlation coefficients between the KR21 coefficients and the criterion-referenced coefficients and PPM correlations between the test-retest and criterion-referenced coefficients. (see TABLE 7)

With regard to the KR21 coefficients, the only relationship which is consistent across cut-off scores is between the KR21 and the k(H) coefficients. The correlations are positive for each test length and cut-off score and are significant in 9 out of 12 cases.

The relationship between the KR21 and pcs(s) coefficients of agreement is erratic, inverse for all test lengths at the 60 percent cut-off score, mixed at the 80 percent cut-off score and positive for all test lengths at the 100 percent cut-off score.

The relationship between the KR21 coefficients and k is also erratic, positive, and inverse at the same cut-off score, depending on test length.

These results seem to indicate that the KR21 coefficient and the criterion-referenced coefficients are assessing different types of reliability. The KR21 coefficients are a measure of the internal consistency of the tests while the criterion-referenced coefficients assess the consistency of

mastery over two testings. Thus, although the two types of coefficients may be highly related they cannot be used interchangeably.

To some extent this analysis was limited by the small number of tests available for the analysis. However, the lack of a pattern of association between the KR21 and the criterion-referenced coefficients does seem to indicate that these norm-referenced and criterion-referenced indices are not assessing the same type of reliability.

The analysis of the relationship between the TRT and the criterion-referenced coefficients reveals no pattern between any of the two types of coefficients. Those relationships which do exist are for the shorter test lengths. Since the criterion-referenced indices are concerned with assessing consistency of mastery from one test administration to another (k) or the estimate thereof (k(H), pc(s)(s), pc(s)(c), there would be expected a higher degree of relationship between them and the test-retest coefficients. Perhaps the limited number of available tests, particularly the 8- and 9-item tests, resulted in correlations which were not very meaningful. (see TABLE 8)

The mastery level may be a crucial difference in that a change of one point from pre-test to post-test might not affect the TRT coefficient much while if the one point change were the difference between mastery and non-mastery, it would affect the criterion-referenced coefficient. Thus the

relationship between the TRT and the criterion-referenced coefficients would not be very strong.

In conclusion, each of the three reliability coefficients examined has been found to have somewhat different properties although similarities do exist.

Coefficient k which can only be calculated when there are two test administrations generally has the lowest values and the largest standard errors. The size of k increases with increased test length, decreases when a cut-off score is set at an extreme point such as 100 percent, and seems to be maximized by larger sample sizes which also results in smaller standard errors. The heterogeneity of test content had a mixed effect on k, but it is suspected that truly heterogeneous test items might cause k to have lower values particularly if the item difficulties were different.

The estimate of kappa, k(H), generally had slightly larger mean values than the mean values of k and smaller standard errors. The mean values of k(H) increased with test length, and the means values were highest for a cut-off score of 80 percent. The change in sample size did not predictably affect k(H) were highest for the low ability groups. The violation of test item homogeneity appeared to reduce k(H).

The coefficient of agreement pc(s) from both the simple and the compound binomial models was the highest reliabilty coefficient with the smallest standard errors, across all

analyses. Part of the reason for the high values of pc(s) is the difference in scaling between pc(s) and the kappa coefficients. The mean values of pc(s) did increase with test length and with respect to cut-off score, the maximum mean values occurred at the 60 percent cut-off score. The size of the pc(s) coefficient varied for ability level by cut-off score and test length, being higher for the high ability groups at the 100 percent cut-off score. The pc(s) coefficients' mean values were higher for one classroom samples than for the larger group, and the heterogeneity of test content increased the size of the mean values of pc(s).

The use of k(H) as an estimate of k can be reommended only under certain conditions. The difference score analyses generally revealed that k(H) overestimated k for all conditions except heterogeneous test content. k(H) appears to be the best estimate of k for the following conditions: for short tests; for homogeneous ability groups; for larger sample sizes with a cut-off score in the middle of the test score distribution; and for heterogeneous test content.

While with two test administrations one could compute the standard k, the size of the standard errors of k are so large that it is questionable how much information one is getting. k(H) has been shown to consistently overestimate k somewhat, but the standard errors for k(H) are very small. Thus, in some instances one might choose to administer a test only once and

then calculate k(H), knowing that it will usually be an overestimate, but that the standard errors are small. Another possibility is that k, the standard, may be an underestimate due to performance factors which cannot be controlled. In this case k(H) might then be a more accurate estimate of reliability. In either case it seems reasonable to regard k(H) as an upper bound to reliability. Thus the calculation of a low value for k(H) should be regarded as a signal of questionable reliability for a test. With the appropriate testing situation the use of k(H) as an estimate could be recommended instead of scheduling a double test administration in order to calculate k. Thus the decision to use one of these coefficients must be made by considering the length and composition of the test, the cut-off score used for mastery, and the size of the sample. The intent of this research has been to provide information about the behavior of these three reliability coefficients based on actual criterion-referenced test data which were used to make decisions about students' progress in a mathematics program in a public school system. As such many of the findings should be generalizable to similar situations.

TABLE 1

The effect of test length (N) on the expected value and s.e. of the three reliability coefficients

Test Length	E(k)	s.e. k	$E(\hat{k}_H)$	s.e. $\hat{k}_H$	$E(p_{cs})$ (simple)	s.e. p_{cs} (simple)	$E(p_{cs})$ (compound)	s.e. p_{cs} (compound)
5 items	.375	.091	.513	.057	.709	.039	.888	.042
7 items	.388	.078	.505	.032	.699	.022	.877	.024
8 items	.511	-	.559	-	.754	-	.890	-
9 items	.335	-	.429	-	.724	-	.956	-
10 items	.450	-	.725	-	.826	-	.958	-
13 items*	.354	-	.796	-	.802	-	.954	-
Average	.402	.085	.588	.045	.752	.031	.921	.033

* - Not mean values because only one group took this test length.

TABLE 2

The effect of cut-off score and test length on the expected value and s.e. of the three reliability coefficients

Cut-off Score	Test Length	E(k)	s.e. k	$E(\hat{k}_H)$	s.e. $\hat{k}_H$	$E(p_{cs})$ (simple)	s.e. p_{cs} (simple)	$E(p_{cs})$ (compound)	s.e. p_{cs} (compound)
60%	5 items	.397	.182	.433	.176	.830	.051	.945	.053
	7 items	.411	.195	.482	.178	.863	.046	.964	.038
	8 items	.487	.085	.552	.117	.817	.023	.938	.046
	9 items	.292	.118	.405	.106	.832	.044	.982	.012
	10 items	.601		.756		.903		.971	
	13 items*	.703		.809		.927		.995	
	Average	.481	.145	.572	.144	.862	.041	.968	.037
80%	5 items	.375	.097	.456	.153	.710	.053	.888	.064
	7 items	.388	.068	.505	.140	.699	.068	.877	.058
	8 items	.511	.107	.559	.098	.754	.023	.890	.046
	9 items	.336	.108	.430	.092	.724	.035	.956	.028
	10 items	.497		.760		.826		.958	
	13 items*	.354		.796		.802		.954	
	Average	.410	.095	.584	.121	.753	.045	.921	.049
100%	5 items	.269	.049	.416	.137	.780	.081	.743	.073
	7 items	.259	.086	.443	.132	.733	.073	.763	.070
	8 items	.193	.181	.458	.068	.795	.071	.778	.077
	9 items	.333	.120	.354	.078	.730	.049	.750	.027
	10 items	.196		.670		.788		.890	
	13 items*	.264		.716		.806		.892	
	Average	.252	.109	.509	.104	.772	.069	.802	.062

* - Not mean values because only one group took this test length.

TABLE 3

The combined effect of ability level and cut-off score on the expected value and s.e. of the three reliability coefficients for low ability and high ability students

Cut-off Score	Ability Level	E(k)	s.e.(k)	E($\hat{k}$)	s.e.($\hat{k}$)	E(p_{cs}) (simple)	s.e.(p_{cs}) (simple)	E(p_{cs}) (compound)	s.e.(p_{cs}) (compound)
60%	Low	.473	.121	.617	.041	.815	.026	.936	.028
	High	.346	.096	.471	.046	.845	.023	.967	.017
80%	Low	.487	.105	.608	.042	.779	.028	.885	.029
	High	.381	.093	.494	.043	.726	.027	.905	.032
100%	Low	.134	.199	.509	.047	.822	.020	.850	.031
	High	.163	.109	.424	.046	.718	.022	.757	.033

TABLE 4

The effect of sample size on the expected value and s.e. of the three reliability coefficients by cut-off score for one classroom (N = 25) and two classroom (N = 50) samples

Cut-off Score	Sample Size	E(k)	s.e.(k)	E($\hat{k}$)	s.e.($\hat{k}$)	E(p_{cs}) (simple)	s.e.(p_{cs}) (simple)	E(p_{cs}) (compound)	s.e.(p_{cs}) (compound)
60%	One class	.425	.140	.533	.050	.894	.019	.984	.010
	Two classes	.511	.081	.551	.066	.879	.019	.975	.015
80%	One class	.378	.105	.548	.052	.794	.029	.954	.004
	Two classes	.385	.072	.565	.038	.768	.022	.940	.026
100%	One class	.064	.117	.495	.056	.734	.020	.803	.038
	Two classes	.242	.089	.501	.040	.739	.015	.803	.036

TABLE 5

The effect of test length on the expected value and s.e. of the reliability coefficients for tests of homogeneous and heterogeneous test content

Test Length	Test Content	E(k)	s.e.(k)	E($\hat{k}$)	s.e.($\hat{k}$)	E(p_{cs}) (simple)	s.e.(p_{cs}) (simple)	E(p_{cs}) (compound)	s.e.(p_{cs}) (compound)
5 items	Homogeneous	.375	.091	.513	.057	.709	.039	.888	.042
	Heterogeneous	.295	.154	.260	.119	.789	.092	.976	.028
7 items	Homogeneous	.388	.078	.505	.032	.699	.022	.877	.024
	Heterogeneous	.323	.136	.296	.102	.744	.087	.969	.022
8 items	Homogeneous	.511	-	.559	-	.754	-	.890	-
	Heterogeneous	.230	.194	.157	.104	.854	.132	.997	.006
9 items	Homogeneous	.335	-	.429	-	.724	-	.956	-
	Heterogeneous	.423	.115	.415	.062	.795	.044	.975	.014
10 items	Homogeneous	.450	-	.725	-	.826	-	.958	-
	Heterogeneous	.570	.102	.566	.082	.761	.037	.937	.019
13 items	Homogeneous	.354	-	.796	-	.802	-	.954	-
	Heterogeneous	.520	.081	.622	.052	.762	.024	.967	.014

TABLE 6

Validity coefficients and kappa coefficients for 4 mastery tests

Grade	Objective	Test Length	kappa Coefficient	Validity Coefficient	KR21
5	Multiplication 07-K	5 items	.397	.343	.731
7	Multiplication 08-L	5 items	.216	.119	.385
5	Division 08-J	5 items	.450	.287	.719
7	Division 10-N	5 items	.214	.438	.519

TABLE 7

PPM correlations between the KR21 and each criterion-referenced coefficient by test length and cut-off score

Cut-Off Score	Test Length	PPM between KR21 and k	PPM between KR21 and $\hat{k}_H$	PPM between KR21 and $p_{cs}(s)$	PPM between KR21 and $p_{cs}(c)$
60%	5 items	.566	.753**	-.396	-.103
	7 items	-.213	.969**	-.698*	-.550
	8 items	-.524	.995***	-.930*	-.812
	9 items	-.518	.977*	-.660	-.984*
80%	5 items	.745**	.637*	.432	-.044
	7 items	.654*	.902***	.824**	-.076
	8 items	.792	1.00***	.823	-.691
	9 items	.475	.934	-.134	-.684
100%	5 items	.135	.475	.420	.085
	7 items	-.758**	.946***	.831**	.916***
	8 items	.030	.981**	.980**	.992***
	9 items	-.898	.865	.756	.821

* $p < .05$
** $p < .01$
*** $p < .001$

TABLE 8

PPM correlations between the TRT and each criterion-referenced coefficient by test length and cut-off score

Cut-Off Score	Test Length	PPM between TRT and k	PPM between TRT and $\hat{k}_H$	PPM between TRT and $p_{cs}(s)$	PPM between TRT and $p_{cs}(c)$
60%	5 items	.739**	.693*	-.727*	-.497
	7 items	.630*	.265	-.281	-.356
	8 items	-.390	-.095	-.222	-.237
	9 items	.349	-.909	.790	+.994**
80%	5 items	.761**	.706*	.859***	.501
	7 items	.249	.082	.007	-.209
	8 items	.043	-.102	-.742	.136
	9 items	-.677	-.904	.319	.890
100%	5 items	-.174	.606*	.833**	.131
	7 items	.063	.144	.405	.120
	8 items	.950*	-.015	-.299	-.550
	9 items	.811	-.741	-.870	-.690

*p < .05
**p < .01
***p < .001

Berk, Ronald A. A Consumers' Guide to Criterion-Referenced Test Reliability, Journal of Educational Measurement, 1980, 17(4), 323-349.

Cohen, J.S. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Hambleton, R. K., Swaminathan, H., Algina, J., and Coulson, D.B. Criterion-referenced testing and measurement. A review of technical issues and development. Review of Educational Research, Winter 1978, 48, 1-47.

Huynh, Huynh. On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement. 1976, 13, 253-264.

Subkoviak, M. Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 1976, 13, 265-275.

Swaminathan, H., Hambleton, R. K., and Algina, J., Reliability of criterion-referenced tests: A decision-theoretic formulation. Journal of Educational Measurement, 1974, 11, 263-268.